Attempting to Bypass Alignment from Comparable Corpora via Pivot Language

Alexis Linard **Béatrice Daille** **Emmanuel Morin**
Université de Nantes, LINA UMR CNRS 6241
2 rue de la Houssinière, BP 92208
44322 Nantes Cedex 03, France
firstname.lastname@univ-nantes.fr

Abstract

Alignment from comparable corpora usually involves two languages, one source and one target language. Previous works on bilingual lexicon extraction from parallel corpora demonstrated that more than two languages can be useful to improve the alignments. Our works have investigated to which extent a third language could be interesting to bypass the original alignment. We have defined two original alignment approaches involving pivot languages and we have evaluated over four languages and two pivot languages in particular. The experiments have shown that in some cases the quality of the extracted lexicon has been enhanced.

1 Introduction

The main goal of this work is to investigate to which extent bilingual lexicon extraction using comparable corpora can be improved using a third language when dealing with poor resource language pairs. Indeed, the quality of the result of the extracted bilingual lexicon strongly depends on the quality of the resources, that is to say the corpora and a general language bilingual dictionary. In this study, we stress the key role of the potential high quality resources of the pivot language (Chiao and Zweigenbaum, 2004; Morin and Prochasson, 2011; Hazem and Morin, 2012). The idea of involving a third language is to benefit from the lexical information conveyed by the additional language. We also assume that in the case of not so usual language pairs the two comparable corpora are of medium quality, and the bilingual dictionary seems weak, due to the nonexistence of such a dictionary. We expect as a consequence a bad quality of the extracted lexicon. Nevertheless, we are highly confident that a language for which we have of a lot of resources can thwart the effect of the poor original resources. English is probably the first language in term of work and resources in Natural Language Processing, hence it can appear as a good candidate as pivot language.

The paper is organized as follows: we give a short overview of bilingual lexicon extraction standard method in Section 2. Our proposed approaches are described in Section 3. The resources we have used are presented in Section 4 and experimental results in Section 5. Finally, we expose further works and improvements in Sections 6 and 7.

2 Bilingual Lexicon Extraction

Initially designed for parallel corpora (Chen, 1993), and due to the scarcity of this kind of resources (Martin et al., 2005), bilingual lexicon extraction then tried to deal with comparable corpora instead (Fung, 1995; Rapp, 1995). An algorithm using comparable corpora is the standard method (Fung and McKeown, 1997) closely based on the notion of context vectors. Many implementations have been designed in order to do so (Rapp, 1999; Chiao and Zweigenbaum, 2002; Morin et al., 2010). A context vector $\mathbf{w}$ is, for a given word w, the representation of its contexts $ct_1 \dots ct_i$ and the number of occurrences found in the window of a corpus. In this approach, context vectors are calculated both in source and target languages corpora. They are also normalized according to association scores. Then, thanks to a seed dictionary, source context vectors are transferred into target language. The similarity between the translated context vector $\overline{\mathbf{w}}$ for a given source word w to translate and all target context vectors $\mathbf{t}$ lead to the creation of a list of ranked candidate translations. The rank is function of the similarity between context vectors so that the closer they are, the better the ranked translation is.

Research in this field aims at improving the

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 32–37,
Beijing, China, July 30, 2015.

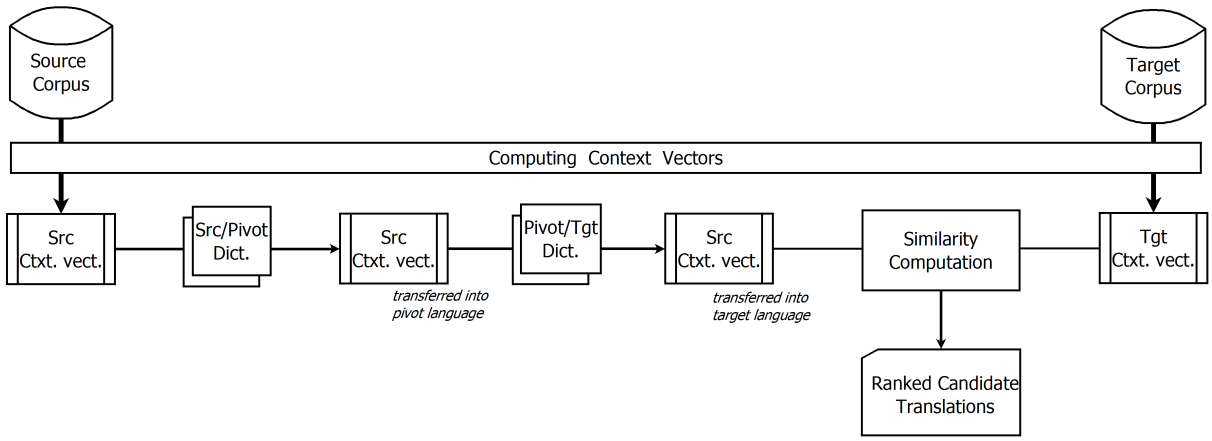

Figure 1: Transferring Context Vectors Successively.

quality of the extracted lexicon. For instance, we can cite the use of a bilingual thesaurus (Déjean et al., 2002), implication of predictive methods for word co-occurrence counts (Hazem and Morin, 2013) or the use of unbalanced corpora (Morin and Hazem, 2014). Among them, and in the case of comparable corpora, we can denote that none looked into pivot-language approaches.

Nevertheless, the idea of involving a pivot language for translation tasks is not recent. Bilingual lexicon extraction from parallel corpora has already been improved via the use of an intermediary language (Kwon et al., 2013; Seo et al., 2014; Kim et al., 2015), so does statistical translation (Simard, 1999; Och and Ney, 2001). Those works lay on the assumption that another language brings additional information (Dagan and Itai, 1991).

3 Alignment Approaches with Pivot Language

In this paper, we present two original approaches which derive from the standard method and involve a third language. We assume that the bilingual dictionary is unavailable or of low quality, but that the source/pivot and pivot/target dictionaries are much better.

3.1 Transferring Context Vectors Successively

The first method, and the most naive is to translate context vectors successively, to start with from source to pivot language, and to follow from pivot to target language. Hence, the context vectors in the source language are computed as it is usually done in the standard method. Then, the second step is to transfer them into the pivot language thanks to a source/pivot dictionary. This operation is done a second time from pivot to target language with a pivot/target dictionary in order to obtain source context vectors translated into target language. We can say that we transferred the context vectors *via* a pivot language. Finally, the last step of similarity computation stays unchanged: for one source word w for which we want to find the translation in target language, we compute the similarity between its context vector transferred successively $\overline{\overline{\mathbf{w}}}$ and all target context vectors $\mathbf{t}$. This method is presented in Figure 1.

3.2 Transposing Context Vectors to Pivot Language

The second method based on pivot dictionaries consists in translating both source and target context vectors into pivot language. Thus, the operation of computing similarity occurs in the vectorial space of the pivot language. In order to do so, the context vector of a word in source language to translate is computed as it is usually done in the standard method. The second step is to transfer the source and target context vectors into the pivot language using source/pivot and target/pivot dictionaries. At this stage, we gather in the pivot language the translated source and all target context vectors. The next and last operation is to compute the similarity between the source context vector transferred into pivot language $\overline{\mathbf{w}}$ and all target context vectors transferred into pivot language $\overline{\mathbf{t}}$. This method is presented in Figure 2.

4 Multilingual Resources

In this paper, we perform translation-candidate extraction from all pairs of languages from/to En-

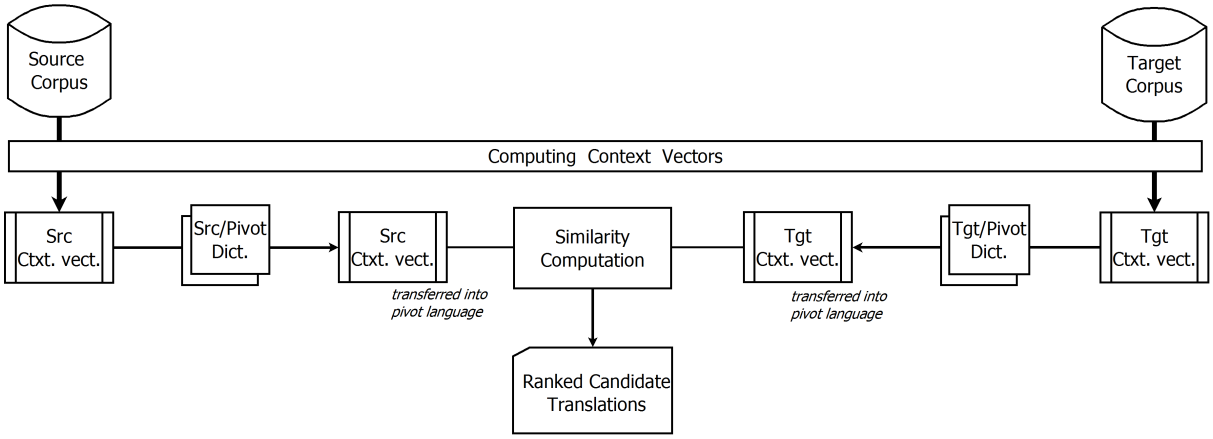

Figure 2: Transposing Context Vectors to Pivot Language.

glish, French, German and Spanish and involving English or French as the pivot language. The use of those pivot languages in particular is motivated by two factors: first, English, because it is the language *by default* we have of in a quasi infinite amount of data, and last, French, because we know that our resources (corpus and dictionaries) are of good quality.

4.1 Comparable Corpora

The first comparable corpus we used during our experiments is the *Wind Energy corpus*[1]. It was built from a crawl of webpages using many keywords related to the wind energy field. The comparable corpus is composed of documents in 7 languages, among others German, English, Spanish and French. The second comparable corpus we used is the *Mobile Technologies corpus*. It was also built by crawling the web. Both of them were composed of 300k to 470k words in each language.

4.2 Bilingual Dictionaries

	EN-DE DE-EN	EN-ES ES-EN	EN-FR FR-EN	FR-ES ES-FR	FR-DE DE-FR	DE-ES ES-DE
#entr.	600k	26k	240k	100k	170k	15k

Table 1: Number of entries in each dictionary.

In order to perform bilingual lexicon extraction from comparable corpora, a bilingual dictionary was mandatory. Nevertheless, we only have of French/English, French/Spanish and French/German dictionaries from the ELRA catalogue[2]. These dictionaries were generalist, and contained few terms related to the Wind Energy and Mobile Technologies domains. So, the French/English, French/Spanish and French/German were reversed to obtain English/French, Spanish/French and German/French dictionaries. As for the others, they were built by triangulation from the ones above (see Table 1). As a consequence, we expect those dictionaries to be very mediocre.

4.3 Reference Lists

	EN	FR	ES	DE
WE	48	58	55	55
MT	52	58	60	88

Table 2: Number of SWT in reference lists.

In order to evaluate the output of the different approaches, terminology reference lists were built from each corpus in each language (Loginova et al., 2012). Depending on the corpus and the language, the lists were composed of 48 to 88 single word terms (abbreviated SWT – see Table 2).

5 Experiments and Results

Pre-processing French, English, Spanish and German documents were pre-processed using TTC TermSuite (Rocheteau and Daille, 2011)[3]. The operations done during pre-processing were the following: tokenization, part-of-speech tagging and lemmatization. Moreover, function words and hapaxes had been removed.

[1] http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html

[2] http://catalog.elra.info/

[3] https://logiciels.lina.univ-nantes.fr/redmine/projects

		Wind Energy					Mobile Technologies				
Lang.	Pivot	Std.	P_1	P_2	R_{MAX}	C	Std.	P_1	P_2	R_{MAX}	C
EN-ES	FR	0.268	**0.390**	**0.374**	*0.646*	65.76%	0.445	**0.523**	**0.467**	*0.882*	66.52%
ES-EN	FR	0.119	**0.232**	**0.233**	*0.491*		0.193	**0.272**	**0.321**	*0.533*	
EN-DE	FR	0.158	0.125	**0.215**	*0.458*	66.21%	0.622	0.205	0.570	*0.896*	68.95%
DE-EN	FR	0.018	0.018	0.018	*0.200*		0.074	0.070	0.069	*0.455*	
FR-DE	EN	0.056	**0.118**	**0.132**	*0.418*	77.63%	0.053	**0.063**	**0.061**	*0.597*	80.06%
DE-FR	EN	0.038	0.028	0.028	*0.151*		0.034	0.023	0.026	*0.432*	
FR-ES	EN	0.366	0.150	0.176	*0.528*	82.36%	0.514	0.275	0.280	*0.807*	82.02%
ES-FR	EN	0.210	0.103	0.117	*0.357*		0.238	0.207	0.186	*0.552*	
ES-DE	FR	0.000	**0.041**	**0.097**	*0.273*	44.24%	0.001	**0.058**	**0.067**	*0.500*	44.02%
	EN	0.000	**0.045**	**0.027**			0.001	**0.033**	**0.035**		
DE-ES	FR	0.001	**0.018**	**0.018**	*0.218*		0.126	**0.355**	**0.347**	*0.585*	
	EN	0.001	**0.018**	**0.018**			0.126	**0.189**	**0.179**		

Table 3: MRR achieved for pivot dictionary based approaches.

Context vectors In order to compute and normalize context vectors, the value $a(ct)$ associated to each co-occurrence ct of a given word w in the corpus was computed. Such value could be computed thanks to Log Likelihood (Fano and Hawkins, 1961) or Mutual Information (Dunning, 1993) for instance. Among them we chose Log Likelihood as its representativity is the most accurate (Bordag, 2008). Context vectors were computed by TermSuite, as one of its components performed this operation.

Similarity measures The so-called similarity could be computed according to Cosine similarity (Salton and Lesk, 1968) or Weighted Jaccard Distance (Grefenstette, 1994). We decided to only present the results achieved using Cosine similarity. The differences between them in term of Mean Reciprocal Rank (MRR) were insignificant.

$$\text{Cosine } (\overline{\mathbf{w}}, \mathbf{t}) = \frac{\sum_k a(\overline{\mathbf{w}}_k) a(\mathbf{t_k})}{\sqrt{\sum_k a(\overline{\mathbf{w}}_k)^2} \sqrt{\sum_k a(\mathbf{t_k})^2}}$$

Evaluation metrics In order to evaluate our approaches, we used Mean Reciprocal Rank (Voorhees, 1999). The strength of this metric is that it takes into account the rank of the candidate translations. Hereinafter, the MRR defined as follows (t stands for the terms to evaluate and r_t for the rank achieved by the system for the good translation of t):

$$MRR = \frac{1}{|t|} \times \sum_{k=1}^{|t|} (\frac{1}{r_{t_k}})$$

Results The MRR achieved for both approaches is shown in Table 3 for Wind Energy and Mobile Technologies corpora respectively. We present, for the sake of comparison, the results achieved by the standard method (Std.), method transferring context vectors successively (P_1) and the method transposing context vectors to pivot language (P_2). We also give additional information, such as the best achievable result according to the reference lists and the words belonging to the filtered corpus (R_{MAX}) and corpora comparability C (Li and Gaussier, 2010).

The corpus comparability metric consists in the expectation of finding the translation in target language for each source word in the corpus. Therewith, it is a good way of measuring the distributional symmetry between two corpora and given a dictionary. We can also notice that the Maximun Recall R_{MAX} is quite low for some pairs of languages: this is due to the high number of hapaxes belonging to the reference lists that were filtered out during pre-processing.

According to the results, we can see that there is a strong correlation between the improvements achieved by pivot based approaches and corpus comparability. We have improved the quality of the extracted bilingual lexicon only in the case of poorly comparable corpora, respectively $\leq$ 65.76% and $\leq$ 66.52% for Wind Energy and Mobile Technologies corpora. For instance, we have increased the MRR from 0.268 to 0.390 and 0.374 in the case of translation from English to Spanish for the Wind Energy corpus, and from 0.126 to 0.355 and 0.347 for German to Spanish via French for the Mobile Technologies corpus.

6 Discussion

In Section 5 we have shown up that results can be enhanced only in the case of poorly comparable pairs of languages. For fairly comparable corpora

	EN-DE DE-EN	EN-ES ES-EN	EN-FR FR-EN	FR-ES ES-FR	FR-DE DE-FR	DE-ES ES-DE
WE	66.21%	65.76%	80.23%	82.36%	77.63%	44.24%
MT	68.95%	66.52%	80.99%	82.02%	80.06%	44.02%

Table 4: Corpora comparability.

($\leq 68\% \leq C \leq 80\%$), results remain unchanged in comparison with the standard approach. Finally, for highly comparable corpora ($C > 80\%$) the quality of the extracted lexicon gets worse.

The interpretation we suggest is the following: given two corpora, S in source language, T in target and a bilingual dictionary source/target $\mathcal{D}$, the comparability is function of S, T, $\mathcal{D}_{S/T}$. Therefore, a low comparability measure can be due to a poor expectation of finding the translation in target language for each source word in the corpus because the two corpora are not lexically close enough, or because the dictionary is weak. We checked this second option, and this is how we substantiate the pivot dictionary based approaches. Thus, the use of source/pivot $\mathcal{D}_{S/P}$ and pivot/target $\mathcal{D}_{P/T}$ dictionary can artificially improve the comparability and enhance the extracted lexicon. We have also remarked that the coverage of dictionaries is an important factor: a large dictionary is better than a shorter.

Of course, we do not pretend that our methods can compare with an initially very highly comparable corpora since the use of pivot dictionaries will introduce more noise than it will bring additional information.

7 Conclusion

We have presented two pivot based approaches for bilingual lexicon extraction from comparable specialized corpora. Both of them lay on pivot dictionaries. We have shown that the bilingual lexicon extraction depends on the quality of the resources. Furthermore, we have also demonstrated that the problem can be fixed involving a third strongly supported language such as English for instance. We have also carried out that the enhancements are function of the comparability of the corpora. These first experiments have shown that using a pivot language can make improvements in the case of poorly comparable initial corpora.

In future works, we will try to benefit from the information brought by an unbalanced pivot corpus. Unlike this article in which we have only looked into pivot dictionaries in order to increase the comparability of the source and target corpora, we think that the next step is to reshape context vectors with a pivot corpus. In addition, we will see whether linear regression models to reshape context vectors can make improvements or not.

Acknowledgments

This work is supported by the French National Research Agency under grant ANR-12-CORD-0020.

References

Stefan Bordag. 2008. A comparison of co-occurrence and similarity measures as simulations of context. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 52–63. Haifa, Israel.

Stanley F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, USA.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–5, Taipei, Taiwan.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2004. Aligning words in french-english non-parallel medical texts: Effect of term frequency distributions. In *Medinfo 2004: Proceedings of the 11th World Congress on Medical Informatics*, pages 23–27, Amsterdam, Netherlands. Ios Pr Inc.

Ido Dagan and Alon Itai. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 130–137, Berkeley, California, USA.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Taipei, Taiwan.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Robert M Fano and David Hawkins. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, Beijing, China.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pages 173–183, Cambridge, Massachusetts, USA.

Gregory Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Springer Science & Business Media.

Amir Hazem and Emmanuel Morin. 2012. Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, pages 288–292, Istanbul, Turkey.

Amir Hazem and Emmanuel Morin. 2013. Word Co-occurrence Counts Prediction for Bilingual Terminology Extraction from Comparable Corpora. In *6th International Joint Conference on Natural Language Processing.*, pages 1392–1400, Nagoya, Japan.

Jae-Hoon Kim, Hong-Seok Kwon, and Hyeong-Won Seo. 2015. Evaluating a pivot-based approach for bilingual lexicon extraction. *Computational Intelligence and Neuroscience*, 2015.

Hong-seok Kwon, Hyeong-won Seo, and Jae-hoon Kim. 2013. Bilingual lexicon extraction via pivot language and word alignment tool. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 11–15, Sofia, Bulgaria, August.

Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652, Beijing, China.

Elizaveta Loginova, Anita Gojun, Helena Blancafort, Marie Guégan, Tatiana Gornostay, and Ulrich Heid. 2012. Reference lists for the evaluation of term extraction tools. In *Proceedings of the 10th International Congress on Terminology and Knowledge Engineering*, Madrid, Spain.

Joel Martin, R Mihalcca, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of The ACL Workshop on Building and Using Parallel Text*, pages 65–74, Ann Arbor, Michigan, USA.

Emmanuel Morin and Amir Hazem. 2014. Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1284–1293, Baltimore, USA.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 27–34, Portland, Oregon, USA.

Emmanuel Morin, Béatrice Daille, Kyo Kageura, and Koichi Takeuchi. 2010. Brains, not Brawn: The Use of "Smart" Comparable Corpora in Bilingual Terminology Mining. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(1):1–23.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Machine Translation Summit*, pages 253–258, Santiago de Compostela, Spain.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526.

Jérôme Rocheteau and Béatrice Daille. 2011. TTC TermSuite: A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 9–12, Chiang Mai, Thailand.

Gerard Salton and Michael E Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36.

Hyeong-Won Seo, Hong-Seok Kwon, and Jae-Hoon Kim. 2014. Extended pivot-based approach for bilingual lexicon extraction. *Journal of the Korean Society of Marine Engineering*, 38(5):557–565.

Michel Simard. 1999. Text-translation alignment: Three languages are better than two. In *Proceedings of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 2–11, College Park, Maryland, USA.

Ellen M. Voorhees. 1999. The trec-8 question answering track report. In *Proceedings of TREC-8*, volume 99, pages 77–82.